Evaluating the size of the bootstrap method for fund performance evaluation[☆]

Abstract

We investigate the validity and reliability of the bootstrap approach in fund performance evaluation by gauging the size. Monte Carlo simulations suggest that cross-sectional dependence may alter the size of this test and we propose a new panel bootstrap approach.

Keywords: Performance evaluation, Bootstrap, Monte Carlo simulation, Unobservable factors

JEL Classification: C15, G11, G12, G23

1. Introduction

One of the cornerstones of the Market Efficiency Hypothesis (EMH) is the principle that active investors (e.g., fund managers) do not have skills to beat the market in a persistent way. In the last decade, the bootstrap method has become an increasingly popular way to evaluate the performance of mutual funds (e.g., Kosowski et al. (2006); Fama and French (2010)), hedge funds (e.g., Kosowski et al. (2007)), pension funds (e.g., Blake et al. (2013)) and even individual investors (Meyer et al. (2012)). The great appeal of this method comes from its simplicity and its ability to circumvent any ex ante parametric assumption on fund alphas (e.g., Kosowski et al. (2006); Fama and French (2010)). This method allows for the generation of the cross-sectional distribution of fund alphas purely due to sampling variability ("luck"), against which, the cross-section of realized alphas obtained from estimating a benchmark model is compared. A significant difference between them is regarded as evidence of genuine skill.

Given the popularity and the seeming superiority of this approach for separating skill from luck, it is surprising that no rigorous statistics analysis has been conducted to examine whether it can actually lead to correct inferences on managers' skill as researchers presumed. Such analysis is essential to the understanding of the recent literature and the appropriate application of bootstrap in future research. Our paper fills this gap. While it is difficult, if not impossible, to examine the validity and reliability of the bootstrap approach, Monte Carlo simulation appears to be the natural choice for this need.

We explicitly investigate the two potential concerns on the bootstrap method in fund performance evaluation. First, the validity and reliability of this approach hinge on sample variations, i.e., the cross-sectional number of funds and time-series observations in the fund performance evaluation area. If sample variation is not enough, the bootstrap method may inevitably lead to a partial instead of a full picture of the underlying population due to the canonical type I error, which prompts us to gauge the size of the application of this method to fund evalua-

tion. This question roots in the strand of literature about hedge fund evaluation, which usually suffers from the short time span of data (Kosowski et al. (2007)). Moreover, the traditional fund-by-fund bootstrap does not take into account of the cross-sectional dependence, brought by the commonly held assets of the fund managers (e.g., Blake et al. (2014)).

We gauge the size of the fund-by-fund bootstrap performance evaluation method in two scenarios: with and without cross-sectional dependence in fund returns. Without cross-sectional dependence, our Monte Carlo simulations demonstrate that the size of the fund-by-fund bootstrap method approaches the conventional statistical significance level (0.05) even when we use a realistic small number of funds and time-series observations, which means that the fund-by-fund bootstrap method has excellent statistical properties in distinguishing skill from luck if the fund returns are not cross-sectionally dependent. With cross-sectional dependence in fund returns, however, the size of the fund-by-fund bootstrap method becomes much larger than 0.05 at any quantiles including the extreme tails, which means that the statistical inferences of this approach are severely biased towards identifying "skills" of fund managers. The "skills" of fund managers identified by this approach in the previous literature may be spurious and simply due to the cross-sectional dependence in fund returns associated with common asset holding. Although to some extent "luck" has been taken into account in this method, cross-sectional dependence in fund returns has not.

Following the recent development on cross-sectional dependence in econometrics (e.g., Bai (2009); Bai and Li (2014)), we take them into account by extending the traditional bootstrap method to a panel case with interactive effects and unobservable factors. This is very different with the existing literature in which all factors are observable (e.g., Kosowski et al. (2006, 2007); Fama and French (2010); Blake et al. (2013, 2014)). It is evident that not all factors are observable (Harvey and Liu (2017*a*)), given the development in the literature of the earlier CAMP-type single-factor market model to the multi-factor models such as the

Fama-French 3-factor, 4-factor, and 5-factor models. It is promising that more factors may be discovered in the future and we treat them as unobservable now. Indeed, one advantage of our model is that it adopts a "let-data-speak" approach to gauge the number of unobservable factors via a principle component analysis addon (Bai (2009)). Moreover, it is easy to see that the usual fixed effects panel data model (e.g., Blake et al. (2014)) is a special case of our panel data model with interactive effects.

The remainder of the paper proceeds as follows. In section 2, we introduce our panel data model with unobservable interactive effects. Section 3 and 4 present the bootstrap procedure and Monte Carlo simulation, respectively. We omit the classical fund-by-fund bootstrap approach for brevity, as it has been well summarized in the extant literature (e.g., Kosowski et al. (2006, 2007); Fama and French (2010); Blake et al. (2013, 2014)). Section 5 concludes.

2. Panel data model with unobservable interactive effects

We propose the following panel data model with unobservable interactive effects to take into account of the cross–sectional dependence:

$$\begin{aligned} r_{it} &= \alpha_i + \beta_i r_{mt} + \varepsilon_{it}, \text{ for } i = 1, \cdots, N, t = 1, \cdots, T, \\ \varepsilon_{it} &= \lambda_i^\top F_t + e_{it}, \end{aligned} \tag{1}$$

where β_i is fund i's risk loading on the market return r_{mt} and α_i is the abnormal return, which is used to measure the fund performance, $F_t(r \times 1)$ is a vector of unobserved common factors, λ_i contains the factor loadings and e_{it} is an idiosyncratic error term.

As mentioned in Blake et al. (2014), the standard framework has the problem that it is potentially incomplete since it excludes fund-specific variables and other common factors which might influence performance. With our approach, this problem can be well solved, as it

can capture not only the observed factors but also unobserved or hidden factors. Note that r_{mt} can be correlated with λ_i alone or with F_t alone, or can be simultaneously correlated with λ_i and F_t. We know that if this correlation exists, then $E(r_{mt}\varepsilon_{it}) \neq 0$, so the traditional ordinary least squares (OLS) estimators of α_i and β_i will be biased and inconsistent.

It is easy to see that model (1) can be rewritten as

$$r_{it} = \alpha_i + \gamma_i^\top z_t + e_{it}, \tag{2}$$

where $\gamma_i = (\beta_i, \lambda_i^\top)^\top$ and $z_t = (r_{mt}, F_t^\top)^\top$. Here $\gamma_i^\top z_t$ can capture all the possible cross-sectional dependence among the N funds, which includes the cross-sectional dependence resulted from not only the observed factor r_{mt} but also the unobserved term $\lambda_i^\top F_t$.

Obviously, model (2) can be written as

$$u_{it} = r_{it} - \gamma_i^\top z_t, \;\; u_{it} = \alpha_i + e_{it}. \tag{3}$$

Ideally, we estimate α_i by the following two steps.

- First, we extract cross-sectional dependence $\gamma_i^\top z_t$ by principle component estimation. Let $\widehat{\gamma_i^\top z_t}$ denotes the estimated values of $\gamma_i^\top z_t$. Under some regularity conditions, we can show that $\widehat{\gamma_i^\top z_t}$ is a consistent estimator of $\gamma_i^\top z_t$ (see, e.g., Bai (2009)).
- Second, we obtain $\widehat{u}_{it}$ by $\widehat{u}_{it} = r_{it} - \widehat{\gamma_i^\top z_t}$. Then from model (3), it is easy to see that a estimator of α_i will be
$$\widehat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}. \tag{4}$$
Then we obtain $\widehat{e}_{it} = \widehat{u}_{it} - \widehat{\alpha}_i$.

Since the extant literature (e.g., Kosowski et al. (2006, 2007); Fama and French (2010); Blake et al. (2013, 2014)) has unanimously proposed zero skills as their null hypothesis, we, in

this paper, focus on the type I error and the size of the bootstrap approach for performance evaluation, which can help alleviate the probability of falsely refusing the EMH.

3. Panel bootstrap procedure for fund performance evaluation

Consider the hypothesis testing problem

$$H_0 : \alpha^{(q)} = 0 \quad \text{versus} \quad H_1 : \alpha^{(q)} \neq 0, \text{ for } q = 0.01, 0.02, 0.03, 0.04, 0.05, ...$$

where $\alpha^{(q)}$ denotes the q-th quantile of the cross-sectional distribution of α.

We evaluate the p value of the test for each quantile via the following bootstrap.

Step 1: Estimate model (2) and obtain $\widehat{e}_{it}$ and $\widehat{\gamma_i^\top z_t}$ for $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$.

Step 2: For the i–th fund, we sample $\widehat{e}^\star_{it}$ from $\{\widehat{e}_{it}\}_{t=1}^T$, then under the null hypothesis $\alpha_i = 0$, we generate the bootstrap sample by

$$r^\star_{it} = \widehat{\gamma_i^\top z_t} + \widehat{e}^\star_{it}.$$

Step 3: For the bootstrap sample $\{r^\star_{it}\}$, do the estimation in Step 1 to get $\widehat{\alpha}_i^{(1)}$ for $i = 1, 2, \cdots, N$.

Step 4: Repeat Step 2 and Step 3 for B times to obtain $\widehat{\alpha}_i^{(b)}$ for $i = 1, 2, \cdots, N$ and $b = 1, 2, \cdots, B$.

For each given b, we compute the quantiles for $\widehat{\alpha}_i^{(b)}$ for $i = 1, 2, \cdots, N$. Let $\widehat{\alpha}_q^{(b)}$ denote the q-th quantile of the bootstrapped alphas for the b-th bootstrap sample and let $\widehat{\alpha}_q$ denote the q-th quantile of estimated alphas obtained based on the original sample.

Then at the q-th quantile, the p-value based on the estimated alphas is calculated by

$$p_q = \frac{\sum_{b=1}^B I(\widehat{\alpha}_q^{(b)} < \widehat{\alpha}_q)}{B}, \tag{5}$$

where $I(A)$ is an indicator function, which takes value of 1 if A is true and zero otherwise. Note that we could also compute the p-value based on the t statistic ($\widehat{t}_\alpha$) of estimated alphas. In the following section, we report simulation results based on both estimated alphas and $\widehat{t}_\alpha$.

4. Monte Carlo simulations

This section selectively examines and presents the results on the size of our procedure and the benchmark bootstrap procedure of Kosowski et al. (2006). Although we have obtained similar results from other fund-by-fund bootstrap methods suggested in the existing literature such as Kosowski et al. (2006) and Fama and French (2010), we omit them for brevity.

We consider the following data generating processes (DGP):

$$\begin{aligned} r_{it} &= \alpha_i + \beta_i r_{mt} + \varepsilon_{it}, \text{ for } i = 1, \cdots, N, t = 1, \cdots, T, \\ \varepsilon_{it} &= \lambda_i^\top F_t + e_{it}, \end{aligned} \tag{6}$$

where $\alpha_i = 0$ for $i = 1, 2, \cdots, N$, β_i is generated from an uniform distribution over the support $[0.5, 2.0]$, r_{mt} is generated from a normal distribution with mean 0.08 and standard deviation 0.15 denoted as $N(0.08, 0.15^2)$ and e_{it} is generated from a normal distribution with mean 0 and standard deviation 0.08 denoted as $N(0, 0.08^2)$.

We use the following two representative DGPs to gauge the size of our proposed procedure.

DGP1: $\lambda_i^\top F_t = 0$. So there is no cross-sectional dependence in ε_{it} under DGP1.

DGP2: $\lambda_i \sim N(0, 1)$ for $i = 1, 2, \cdots, N$ and $F_t \sim N(0, 0.1^2)$ for $t = 1, 2, \cdots, T$. This means that there exists cross-sectional dependence in ε_{it} under DGP2.

Consulting with the actual sample in literature (e.g., Kosowski et al. (2006, 2007); Fama and French (2010); Blake et al. (2013, 2014)), we selectively present the results for the following combinations of N and T: $\{(N, T) : (200, 100), (200, 200), (200, 400), (400, 200), (600, 200)\}$.

We randomly generate 500 simulations for each combination (N, T), and for each simulation, we compute p-value following the procedure in Section 3 based on 500 randomly generated bootstrap samples. Our results hold when we increase the number of simulations.

We plot the simulated size based on the estimated values (t-statistics) of α obtained from our procedure and Kosowski et al. (2006) in Figure 1 (Figure 3) and Figure 2 (Figure 4) for DGP1 and DGP2, respectively. We use "CY" and "KTWW" to denote our proposed panel bootstrap model and the benchmark model developed by Kosowski et al. (2006), respectively.

According to Figure 1 and Figure 2, without cross-sectional dependence, the simulated size of both Kosowski et al. (2006) and our procedure approaches the nominal size (0.05), which means that both of them have excellent statistical properties in distinguishing skill from luck. However, the simulated size of Kosowski et al. (2006) becomes much larger when cross-sectional dependence exists, while our approach stays close to 0.05, which lends our procedure a big advantage over Kosowski et al. (2006). As seen in Figure 3 and Figure 4, this conclusion holds when we use the estimated t-statistics ($\widehat{t}_{\alpha}$) instead of alphas ($\widehat{\alpha}_i$).

5. Conclusion

We gauge the size of the fund-by-fund bootstrap performance evaluation method in two scenarios: without and with cross-sectional dependence in fund returns. Without cross-sectional dependence, our Monte Carlo simulations demonstrate that the simulated size of the fund-by-fund bootstrap method approaches the nominal size (0.05) even when we use a realistic small number of funds and time-series observations, which means that the fund-by-fund bootstrap method has excellent statistical properties in distinguishing skill from luck if the fund returns are not cross-sectionally dependent.

With cross-sectional dependence in fund returns, however, the simulated size of the fund-by-fund bootstrap method becomes much larger than 0.05 at any quantile including the ex-

treme tails, which means that the statistical inferences of this approach are severely biased towards identifying "skills". The "skills" of fund managers identified by this approach in the previous literature may be spurious and simply due to the cross-sectional dependence in fund returns associated with common asset holding. Although to some extent "luck" has been taken into account in this method, cross-sectional dependence in fund returns is not.

To tackle this problem, we propose a panel data model with unobservable interactive effects, which adopts a "let-data-speak" approach to gauge the number of unobservable factors via a principle component analysis addon. The existing fixed-effects panel model (e.g., Blake et al. (2014)) is a special case of our model. The simulated size of the test of our new model approaches the nominal size (0.05), no matter whether there exists cross-sectional dependence or not. The power of the proposed test procedure is out of the scope of this paper, as it may become difficult to identify α_i in some cases, which we leave as a direction for future research.

We provide researchers and practitioners with guidance in selecting specific bootstrap method which is most appropriate, intuition regarding the possible deficiency of their specifications, as well as insights for improving the existing bootstrap method or generating alternative estimation methods (such as Chen et al. (2017); Ferson and Chen (2017); Harvey and Liu (2017*b*)) in fund evaluation in the future.

References

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Bai, J. and Li, K. (2014), 'Theory and methods of panel data models with interactive effects', *Annals of Statistics* **42**(1), 142–170.

Blake, D., Caulfield, T., Ioannidis, C. and Tonks, I. (2014), 'Improved inference in the evaluation of mutual fund performance using panel bootstrap methods', *Journal of Econometrics* **183**(2), 202–210.

Blake, D., Rossi, A. G., Timmermann, A., Tonks, I. and Wermers, R. (2013), 'Decentralized investment management: Evidence from the pension fund industry', *Journal of Finance* **68**(3), 1133–1178.

Chen, Y., Cliff, M. T. and Zhao, H. (2017), 'Hedge funds: The good, the bad, and the lucky', *Journal of Financial and Quantitative Analysis* **forthcoming**.

Fama, E. F. and French, K. R. (2010), 'Luck versus skill in the cross-section of mutual fund returns', *Journal of Finance* **65**(5), 1915–1947.

Ferson, W. and Chen, Y. (2017), 'How many good and bad fund managers are there, really?', *University of Southern California working paper* .

Harvey, C. R. and Liu, Y. (2017*a*), 'Lucky factors', *SSRN working paper 2528780* .

Harvey, C. R. and Liu, Y. (2017*b*), 'Rethinking performance evaluation', *NBER Working Paper* (w22134).

Kosowski, R., Naik, N. Y. and Teo, M. (2007), 'Do hedge funds deliver alpha? a bayesian and bootstrap analysis', *Journal of Financial Economics* **84**(1), 229–264.

Kosowski, R., Timmermann, A., Wermers, R. and White, H. (2006), 'Can mutual fund stars really pick stocks? new evidence from a bootstrap analysis', *Journal of Finance* **61**(6), 2551–2595.

Meyer, S., Schmoltzi, D., Stammschulte, C., Kaesler, S., Loos, B. and Hackethal, A. (2012), 'Just unlucky?–a bootstrapping simulation to measure skill in individual investors investment performance', *SSRN working paper 2023588* .

Figure 1: $\widehat{\alpha}_i$**-based simulated size for DGP1**
The three plots in the first row are corresponding to T=100, 200 and 400 when N=200, and the three plots in the second row are for N=200, 400 and 600 when T=200.

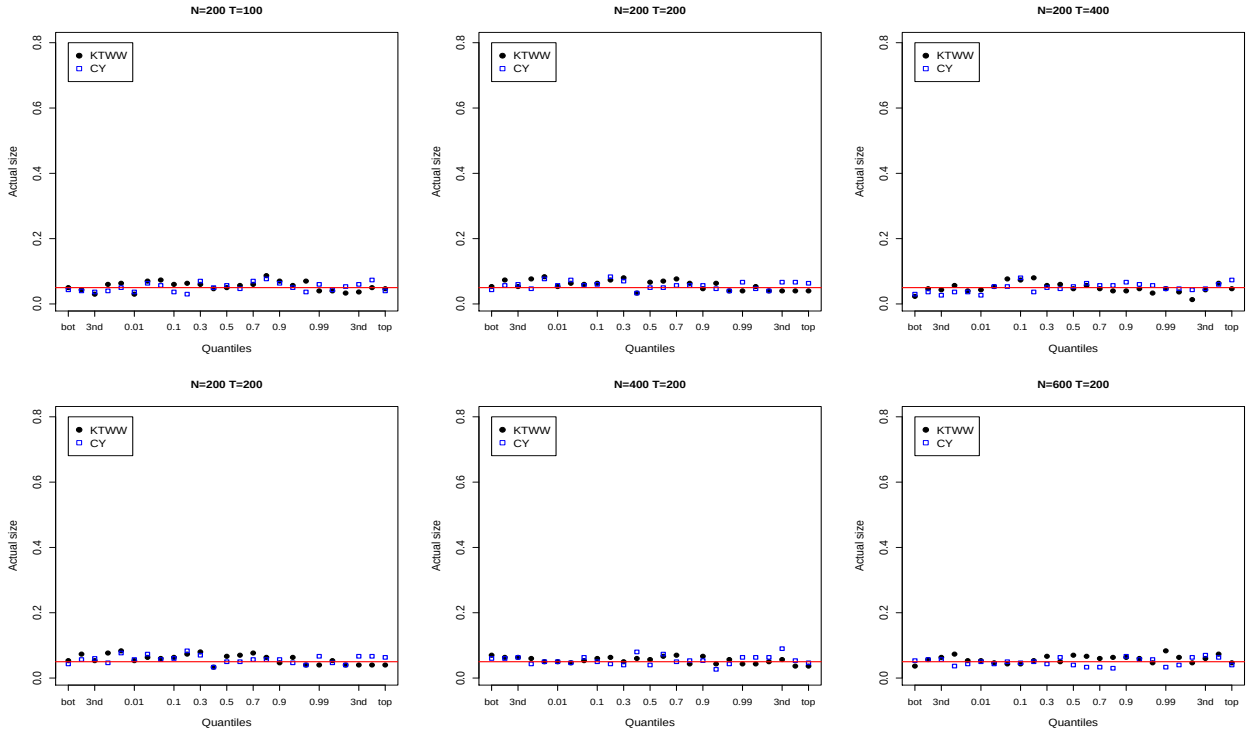

Figure 2: $\widehat{\alpha}_i$**-based simulated size for DGP2**
The three plots in the first row are corresponding to T=100, 200 and 400 when N=200, and the three plots in the second row are for N=200, 400 and 600 when T=200.

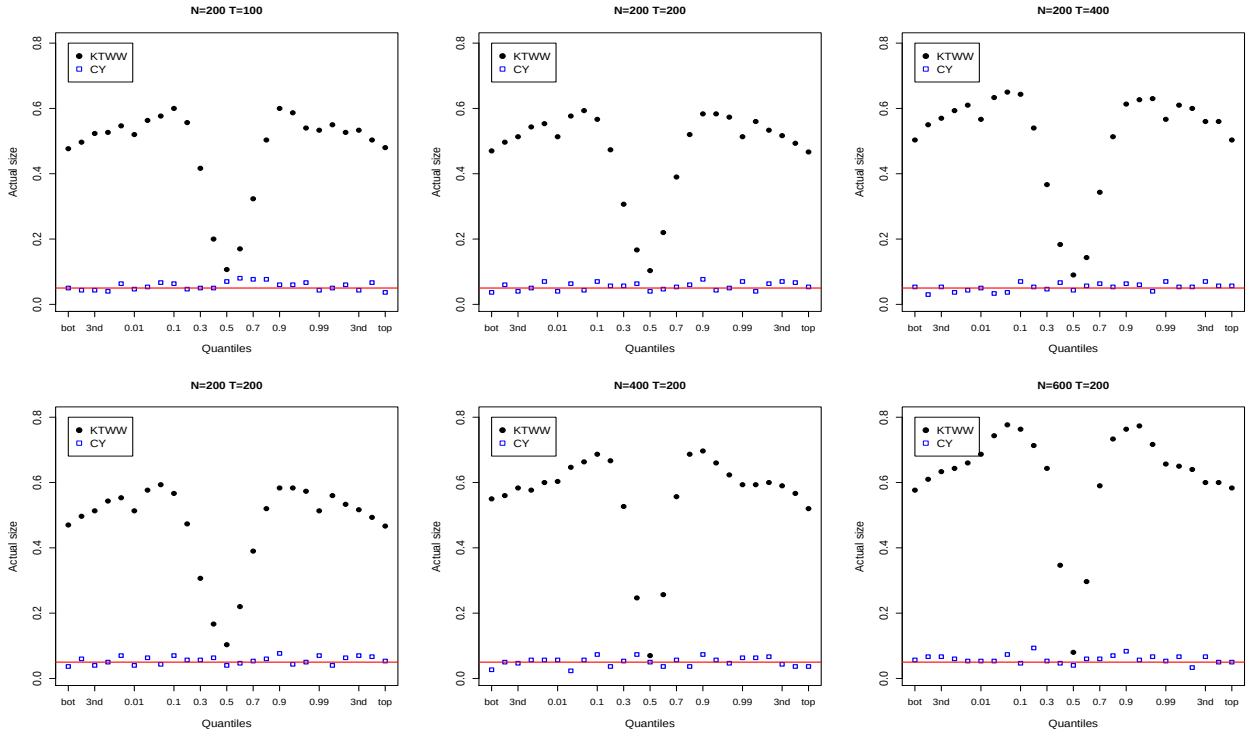

Figure 3: $\widehat{t}_{\alpha}$**-based simulated size for DGP1**
The three plots in the first row are corresponding to T=100, 200 and 400 when N=200, and the three plots in the second row are for N=200, 400 and 600 when T=200.

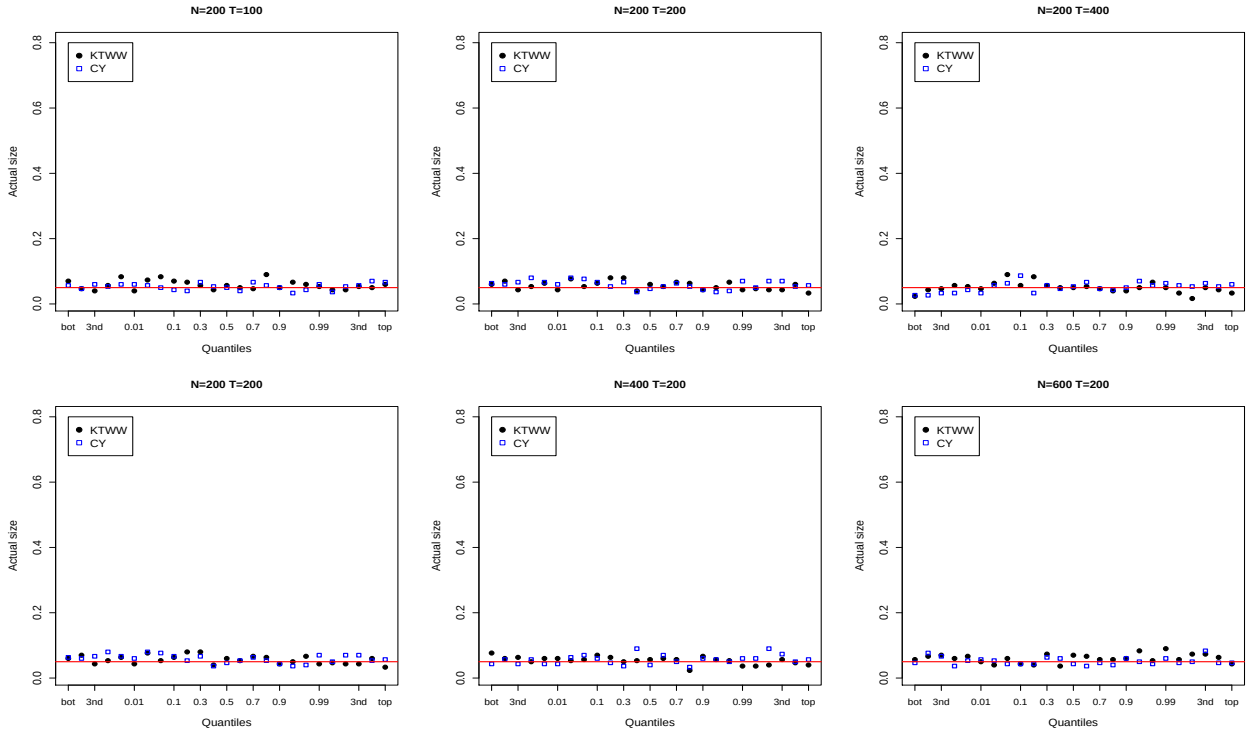

Figure 4: $\widehat{t}_{\alpha}$**-based simulated size for DGP2**
The three plots in the first row are corresponding to T=100, 200 and 400 when N=200, and the three plots in the second row are for N=200, 400 and 600 when T=200.